

3D Reconstruction of Heterogeneous Virus Particles with Statistical Geometric Symmetry

Nan Xu, *Student Member, IEEE*, and Peter C. Doerschuk, *Member, OSA, IEEE*

Abstract—In 3-D reconstruction problems, the image data obtained from cryo electron microscopy is the projection of many heterogeneous instances of the object under study (e.g., a virus). When the object is heterogeneous but has an overall symmetry, it is natural to describe the object as stochastic with symmetrical statistics. This paper presents a maximum likelihood reconstruction approach which allows each object to lack symmetry while constraining the *statistics* of the ensemble of objects to have symmetry. This algorithm is demonstrated on bacteriophage HK97 and is contrasted with an existing algorithm in which each object, while still heterogeneous, has the symmetry. Reconstruction results show that the proposed algorithm eliminates long-standing distortions in previous heterogeneity calculations associated with symmetry axes, and provides estimates that make more biologically sense than the estimates of existing algorithms.

Index Terms—cryo electron microscopy, viruses, symmetrical statistics, maximum likelihood reconstruction, heterogeneity calculation

I. INTRODUCTION

IN virology, discovering the structure of virus particles has been important. The virus particle has a shell of protein, called a “capsid”, which surrounds a cavity containing the viral genome. Typical sizes and molecular weights of the virus particles are 10^2 – 10^3 Å and 10MDa. The capsid is constructed of many repetitions of the same peptide molecule in geometric arrays. The geometry of these spherical viruses is important to their lifecycles. Cryo electron microscopy (cryo EM), which has become an important technique for determining the geometry of a particle, leads to 3-D image reconstruction problems for these biological nanomachines. Specifically, 10^3 – 10^6 virus particles are flash frozen to cryogenic temperatures and imaged. Each image is basically a highly-noisy ($\text{SNR} < 0.1$) 2-D projection of the 3-D electron scattering intensity distribution of the particle. Only one such image can be taken per particle and its projection orientation is unknown. Therefore, the information from many such images is fused to compute a 3-D reconstruction [1].

In the maturation process, the virus particles can be grouped into a number of discrete classes depending on their geometric forms, sizes, pH values, etc. In [2], [3], a reconstruction approach was presented, which includes both discrete classes and continuous variability of the particles within each class. The existing cryo electron microscopy are able to experimentalize

an ensemble of virus particles within one class, which will be the objects to be discussed in the remainder of this paper.

In cryo-electron microscopy, the 3-D structure of each virus particle is described by the 3-D electron scattering intensity function, denoted by $\rho(\cdot)$. The capsid of a virus particle has been well known to obey certain geometric symmetry [4]. The basic and the most popular assumption of the virus particles within one class [5]–[8] is that each instance of the virus particle is identical and exactly obeys the rotational symmetry. In particular,

$$\rho(\mathbf{x}) = \rho_0(\mathbf{x}) \text{ with } \rho_0(\mathbf{x}) = \rho_0(M^{-1}\mathbf{x}), \quad (1)$$

where M is the rotational matrix and $\mathbf{x} \in \mathbb{R}^3$ are real-space coordinates. We call it assumption of “identical individuals”. The reconstruction algorithm based on this rudimentary assumption is usually called the “homogeneous reconstruction with identical individuals” method (HRII). A sufficient assumption introduced in [2], [3] is that instances are different from each other (for example, due to the inherent flexibility of such a large molecular complex) but still each instance exactly obeys the rotational symmetry. In particular,

$$\rho(\mathbf{x}) = \rho(M^{-1}\mathbf{x}). \quad (2)$$

In the remainder of this paper, this assumption is referred as the assumption of “non-identical symmetrical individuals”, and the reconstruction algorithm [2], [3] developed based on this assumption is named as the “heterogeneous reconstruction with symmetrical individuals” (HRSI).

In contrast to this previous work, this paper considers a more realistic and sophisticated view: The different instances of the particle are different and it is the *statistics* of the particle that obey the symmetry not the individual instances. In particular, the mean and covariance of the 3-D structure of the virus particle are invariant under the rotations. This idea is referred as the “symmetrical statistics”.

$$\bar{\rho}(M^{-1}\mathbf{x}) = \bar{\rho}(\mathbf{x}) \text{ and } C_\rho(M^{-1}\mathbf{x}, M^{-1}\mathbf{x}') = C_\rho(\mathbf{x}, \mathbf{x}') \quad (3)$$

where $\bar{\rho}(\mathbf{x}) = \mathbb{E}[\rho(\mathbf{x})]$ and $C(\mathbf{x}, \mathbf{x}') = \mathbb{E}[\rho(\mathbf{x})\rho(\mathbf{x}')]$. We call the new reconstruction algorithm as “heterogeneous reconstruction with symmetrical statistics” (HRSS). Three possible structures described by each of the three assumptions are demonstrated through an example of four fold symmetry in Fig. 1.

The largest class of plant and animal viruses is “spherical” viruses. The capsid of spherical virus particles most common is icosahedral symmetry—the symmetry group of an icosahedron—which is invariant under a total of 60 rotational symmetry operations. In 3-D reality, the structure of a virus

N. Xu is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853 USA e-mail: nx25@cornell.edu

P.C. Doerschuk is with the School of Electrical and Computer Engineering and the Nancy E. and Peter C. Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, 14853 USA e-mail: pd83@cornell.edu

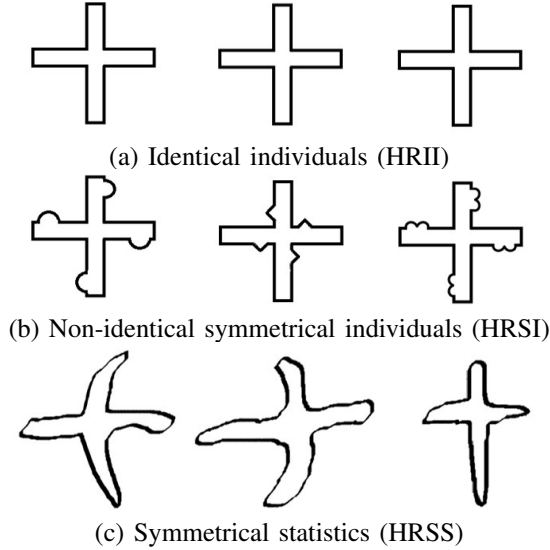


Fig. 1: Four fold symmetry particles based on the three different assumptions and algorithms.

particle can be expressed as a linear combination of basis functions where the weights in the linear combination are Gaussian random variables [2], [3]. In particular,

$$\rho(\mathbf{x}) = \sum_{\zeta} F_{\zeta}(\mathbf{x}) c_{\zeta} \quad (4)$$

where $F_{\zeta}(\mathbf{x})$ is the basis function, c_{ζ} is the weight, and $\mathbf{x} \in \mathbb{R}^3$ are real-space coordinates. Because $\rho(\cdot)$ is real, it is convenient to have real-valued basis functions and weights.

To describe the situations of both “identical individuals” and “non-identical symmetrical individuals”, the complete set of symmetric basis functions of the icosahedral group is sufficient. In the above rudimentary example of 2-D particles with four fold symmetry, these two situations (Fig. 1(a) and Fig. 1(b)) can be both described by the basis functions $e^{i4n\theta} h_l(r)$, where $h_l(r)$ is spherical Bessel functions for degree $l \in \mathbb{N}$. However, to describe the situation in which particles have asymmetric structures but symmetric statistics, a complete orthonormal set of real-valued basis functions with specific rotational properties under the operation of the icosahedral group is required (Section II).

The maximum likelihood (ML) reconstruction algorithm can estimate the mean vector and covariance matrix of the vector of weights c_{ζ} , which is a generalization of classical ML Gaussian mixture parameter estimation [9]. After choosing appropriate basis functions, the estimates of the mean vector and covariance matrix need to be constrained in order to realize the two statistical symmetries as shown in Eq. 3 (Section III). Taking the advantage of the existing reconstruction algorithms, the HRSS algorithm, introduced in this paper, can be developed by adding new basis functions and the additional constraints for the the mean vector and covariance matrix of the weights to the existing HRSI algorithm (Section IV). The reconstruction results are demonstrated in contrast to the previous algorithms in Sections V. As a result, HRSS eliminates the long-standing distortions in heterogeneity cal-

culations associated with symmetry axes introduced by former algorithms. The whole paper is concluded in Section VI.

II. REAL-VALUED BASIS FUNCTIONS

Each real-valued basis function $F_{\zeta}(\mathbf{x})$ is a product of an real-valued angular basis function and a real-valued radial basis function. The real-valued angular basis functions are complete orthonormal square integrable functions on the surface of the sphere, which includes the symmetry properties of the particle. In particular, they are the complete set of real-valued basis functions of the icosahedral group, which were derived and computed in [10] via generalized projection from group theory. Specifically, projection operators are applied to the real-valued spherical harmonics¹ [12, p. 93], which computes a basis function of the icosahedral group as a linear combination of spherical harmonics of a fixed degree l .

A key concept used in this approach is the irreducible representations (irreps) of a finite group, which are sets of 2-D square matrices that are homomorphic under matrix multiplication to the group elements. For the icosahedral group, it is possible to find all irreps matrices that are real-valued, which is required to generate the complete orthonormal basis functions that are real-valued [10]. Suppose that N_{rep} is the number of irreps of the N_g -order group G . Let the set of real-valued matrices in the p th irrep be denoted by $\Gamma^p(g) \in \mathbb{R}^{d_p \times d_p}$ for all $g \in G$ where $p \in \{1, \dots, N_{\text{rep}}\}$. For the icosahedral group, $N_{\text{rep}} = 5$, $N_g = 60$, and $d_p = 1, 3, 3, 4, 5$ for $p = 1, 2, 3, 4, 5$, respectively.

The resulting real-valued angular basis functions associated with the degree l and the p th irrep, denoted by $I_{p;l,n}(\mathbf{x}/x)$ for $n \in \{1, \dots, N_{p;l}\}$, have the following properties.

- 1) Each $I_{p;l,n}$ is a d_p -dimensional real-valued vector function, i.e., $I_{p;l,n} \in \mathbb{R}^{d_p}$.
- 2) The $I_{p;l,n}$ functions are orthonormal on the surface of the sphere;
- 3) The subspace of square integrable functions on the surface of the sphere defined by spherical harmonics of index l , contains a set of $I_{p;l,n}$ functions with a total of $2l + 1$ components.
- 4) Each $I_{p;l,n}$ function has a specific transformation property under rotations from the icosahedral group [12, p. 20], in particular,

$$I_{p;l,n}(R_g^{-1}\mathbf{x}/x) = (\Gamma^p(g))^T I_{p;l,n}(\mathbf{x}/x) \quad (5)$$

where T is transpose not Hermitian transpose and R_g is the 3×3 rotation matrix corresponding to the g th element of the group.

These properties were numerically verified from the close relationship between the $I_{p;l,n}$ and spherical harmonics. Furthermore, it follows that the family of $I_{p;l,n}$ is a complete basis for square integrable functions on the surface of the sphere. Examples of $I_{p;j;l,n}$ functions for $l = 15$ are visualized in Figure 2. Note that $p = 1$ exhibits all of the symmetries of

¹Spherical harmonics are denoted by $Y_{l,m}(\theta, \phi)$ where the degree l satisfies $l \in \mathbb{N}$, the order m satisfies $m \in \{-l, \dots, l\}$ and (θ, ϕ) are the angles of spherical coordinates with $0 \leq \theta \leq \pi$ and $0 \leq \phi \leq 2\pi$ [11, Section 14.30, pp. 378–379].

an icosahedron. The angular basis functions used in the two previous algorithms, HRH and HRSI, are just the angular basis functions at $p = 1$.

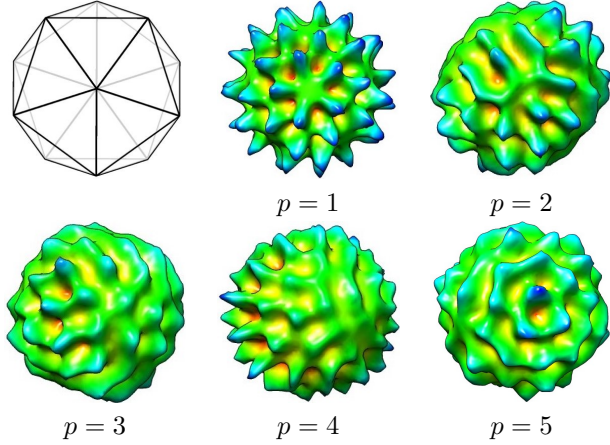


Fig. 2: An icosahedron with one of each type of symmetry axis (2-, 3-, and 5-fold) shown and example angular basis functions with $l = 10$ and $p \in \{1, \dots, N_{\text{rep}}\}$. The surfaces of 3-D objects defined by $\xi(\mathbf{x}) = 1$ for $x \leq \kappa_1 + \kappa_2 I_{p;l,n}(\mathbf{x}/x)$ and 0 otherwise, where κ_1 and κ_2 are chosen so that $\kappa_1 + \kappa_2 I_{p;l,n}(\mathbf{x}/x)$ varies between 0.5 and 1, are visualized by UCSF Chimera where the color indicates the distance from the center of the object.

The radial basis functions $h_{l,q}(x)$ for $q \in \{1, 2, \dots\}$ are exactly the family of Bessel functions used in [13, Section IIIB], which are real-valued and have the orthonormal condition, specifically,

$$\int_{r=0}^{\infty} h_{l,q}(x) h_{l,q'}(x) x^2 dx = \delta_{q,q'}. \quad (6)$$

The reason that $h_{l,q}$ depends on l is that this makes it possible to have symbolic formulas for the 3-D Fourier transform of $\rho(\mathbf{x})$.

Let ζ be the shorthand for l, n, q , which has the total number that associates with the p th irrep of the icosahedral group to be $N_{p;\zeta}$, i.e., $\sum_{\zeta=1}^{N_{p;\zeta}} = \sum_{l=1}^{\infty} \sum_{n=1}^{N_{p;l}} \sum_{q=1}^{\infty}$. Each real-valued basis function, denoted by $F_{p;\zeta}(\mathbf{x})$, is the product of the n th angular basis function and the q th radial basis function that associate with the p th irrep of the icosahedral group and l th degree of spherical harmonics, specifically, $F_{p;\zeta}(\mathbf{x}) = I_{p;l,n}(\mathbf{x}/x) h_{l,q}(x)$. The real-valued basis functions, $F_{p;\zeta}$ for $p \in \{1, \dots, N_{\text{rep}}\}$ and $\zeta \in \{1, \dots, N_{p;\zeta}\}$ have the following properties.

- i) Each $F_{p;\zeta}$ is a d_p -dimensional real-valued vector function, i.e., $F_{p;\zeta} \in \mathbb{R}^{d_p}$.
- ii) The $F_{p;\zeta}$ functions construct a complete orthonormal basis because of the orthonormality of $I_{p;l,n}$ and $h_{l,q}$, specifically,

$$\int_{\mathbf{x}} F_{p;\zeta}(\mathbf{x}) F_{p';\zeta'}^T(\mathbf{x}) d\mathbf{x} = I_{d_p} \delta_{p,p'} \delta_{\zeta,\zeta'}. \quad (7)$$

(Please see Appendix A for the derivation.)

- iii) Each $F_{p;\zeta}$ satisfies the following transformation because of Eq. 5, specifically, for all $g \in G$,

$$\begin{aligned} F_{p;\zeta}(R_g^{-1}\mathbf{x}) &= I_{p;l,n}(R_g^{-1}\mathbf{x}/x) h_{l,q}(x) \\ &= (\Gamma^p(g))^T I_{p;l,n}(\mathbf{x}/x) h_{l,q}(x) \\ &= (\Gamma^p(g))^T F_{p;\zeta}(\mathbf{x}). \end{aligned} \quad (8)$$

III. CONSTRAINTS ON WEIGHTS CAUSED BY SYMMETRICAL STATISTICS

The weight corresponding to the basis function $F_{p;\zeta}$, denoted by $c_{p;\zeta}$, is a real-valued vector, i.e., $c_{p;\zeta} \in \mathbb{R}^{d_p}$. Eq. 4 can be restated as

$$\rho(\mathbf{x}) = \sum_{p=1}^{N_{\text{rep}}} \sum_{\zeta=1}^{N_{p;\zeta}} F_{p;\zeta}^T(\mathbf{x}) c_{p;\zeta}. \quad (9)$$

The index τ in Eq. 4 is the shorthand of p and ζ , which has the total number N_{τ} to satisfy $N_{\tau} = \sum_{p=1}^{N_{\text{rep}}} N_{p;\zeta}$. The vector of weights $\mathbf{c} = [c_1^T, \dots, c_{N_{\tau}}^T]^T = [c_{N_{p=1,\zeta=1}}^T, \dots, c_{N_{p=1,\zeta=N_{p=1,\zeta}}^T}^T, \dots, c_{N_{p=N_{\text{rep}},1}}^T, \dots, c_{N_{p=N_{\text{rep}},\zeta}}^T]^T$ has Gaussian distribution $N(\bar{\mathbf{c}}, V)$, where $\bar{\mathbf{c}} = \mathbb{E}[\mathbf{c}] \in \mathbb{R}^{N_{\tau} \times 1}$ and $V = \mathbb{E}[\mathbf{c}\mathbf{c}^T] \in \mathbb{R}^{N_{\tau} \times N_{\tau}}$.

Our assumption of symmetrical statistics for the spherical virus particles is that the mean and the correlation function of the 3-D scattering intensity are invariant under all 60 rotations of the icosahedral group, which can be translated into the following two conditions, for all $g \in G$,

$$\bar{\rho}(R_g^{-1}\mathbf{x}) = \bar{\rho}(\mathbf{x}), \quad (10)$$

$$C_{\rho}(R_g^{-1}\mathbf{x}, R_g^{-1}\mathbf{x}') = C_{\rho}(\mathbf{x}, \mathbf{x}'), \quad (11)$$

where $\bar{\rho}(\mathbf{x}) = \mathbb{E}[\rho(\mathbf{x})]$, $C_{\rho}(\mathbf{x}, \mathbf{x}') = \mathbb{E}[\rho(\mathbf{x})\rho(\mathbf{x}')]^T$ and R_g is the g th rotation matrix in group G . These two conditions induce constraints on the weights of basis functions, in particular, the mean vector $\bar{\mathbf{c}}$ and covariance matrix V .

A. Constraint caused by the first order statistics

The symmetry on the first order statistics (Eq. 10) induces a constraint on the mean of the weights. By definition (Eq. 9) and the transformation property (Eq. 8), the RHS of Eq. 10 can be expanded as

$$\begin{aligned} \bar{\rho}(\mathbf{x}) &= \mathbb{E} \left[\sum_p \sum_{\zeta} F_{p;\zeta}^T(\mathbf{x}) c_{p;\zeta} \right] \\ &= \sum_p \sum_{\zeta} F_{p;\zeta}^T(\mathbf{x}) \mathbb{E}[c_{p;\zeta}] \end{aligned} \quad (12)$$

and the LHS of Eq. 10 becomes

$$\begin{aligned} \bar{\rho}(R_g^{-1}\mathbf{x}) &= \sum_p \sum_{\zeta} F_{p;\zeta}^T(R_g^{-1}\mathbf{x}) \mathbb{E}[c_{p;\zeta}] \\ &= \sum_p \sum_{\zeta} \left(\Gamma^p(g)^T F_{p;\zeta}(\mathbf{x}) \right)^T \mathbb{E}[c_{p;\zeta}] \\ &= \sum_p \sum_{\zeta} F_{p;\zeta}^T(\mathbf{x}) \Gamma_c^p(g) \mathbb{E}[c_{p;\zeta}], \end{aligned} \quad (13)$$

Equating Eq. 13 and Eq. 12 to give

$$\sum_p \sum_{\zeta} F_{p;\zeta}^T(\mathbf{x}) (\Gamma_c^p(g) - I_d) \mathbb{E}[c_{p;\zeta}] = 0, \quad (14)$$

which holds for all $g \in G$. Because only the $\Gamma_c^{p=1}(g) = 1$ for all $g \in G$ and $\Gamma_c^{p \neq 1}(g)$ for all $g \in G$ are real-valued full matrices, the mean of $c_{p;\zeta}$, needs to satisfy

$$\mathbb{E}[c_{p;\zeta}] = \mathbb{E}[c_{p;\zeta}] \delta_{p,1}. \quad (15)$$

In other words, the entries of \bar{c} that corresponding to the p th ($p \neq 1$) irrep are constrained to be zero.

B. Constraint caused by the second order statistics

The symmetry condition on the second order statistics (Eq. 11) can be achieved by constraints on the covariance matrix of the weights.

Similar to the derivation of Eq. 12-13, the RHS of Eq. 11 can be expressed as

$$\begin{aligned} C_\rho(\mathbf{x}, \mathbf{x}') &= \mathbb{E} \left[\left(\sum_p \sum_\zeta F_{p;\zeta}^T(\mathbf{x}) c_{p;\zeta} \right) \left(\sum_{p'} \sum_{\zeta'} F_{p';\zeta'}^T(\mathbf{x}') c_{p';\zeta'} \right)^T \right] \\ &= \sum_p \sum_\zeta \sum_{p'} \sum_{\zeta'} F_{p;\zeta}^T(\mathbf{x}) \mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] F_{p';\zeta'}(\mathbf{x}'), \end{aligned} \quad (16)$$

and the LHS of Eq. 11 can be expressed as

$$\begin{aligned} C_\rho(R_g^{-1}\mathbf{x}, R_g^{-1}\mathbf{x}') &= \sum_p \sum_\zeta \sum_{p'} \sum_{\zeta'} F_{p;\zeta}(R_g^{-1}\mathbf{x}) \mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] F_{p';\zeta'}(R_g^{-1}\mathbf{x}') \\ &= \sum_p \sum_\zeta \sum_{p'} \sum_{\zeta'} \left((\Gamma_g^p(g))^T F_{p;\zeta}(\mathbf{x}) \right)^T \mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] \left((\Gamma_g^{p'}(g))^T F_{p';\zeta'}(\mathbf{x}') \right) \\ &= \sum_p \sum_\zeta \sum_{p'} \sum_{\zeta'} F_{p;\zeta}^T(\mathbf{x}) \Gamma_g^p(g) \mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] (\Gamma_g^{p'}(g))^T F_{p';\zeta'}(\mathbf{x}') \end{aligned} \quad (17)$$

On both sides of the symmetry condition (Eq. 11), multiply on the left by $F_{p;\zeta}(\mathbf{x})$ and on the right by $F_{p';\zeta'}^T(\mathbf{x}')$ for random (p, ζ) and (p', ζ') , and integrate over the 3D coordinates \mathbf{x} and \mathbf{x}' . Specifically,

$$\begin{aligned} \int_{\mathbf{x}} \int_{\mathbf{x}'} F_{p;\zeta}(\mathbf{x}) C_\rho(\mathbf{x}, \mathbf{x}') F_{p';\zeta'}^T(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \\ = \int_{\mathbf{x}} \int_{\mathbf{x}'} F_{p;\zeta}(\mathbf{x}) C_\rho(R_g^{-1}\mathbf{x}, R_g^{-1}\mathbf{x}') F_{p';\zeta'}^T(\mathbf{x}') d\mathbf{x} d\mathbf{x}'. \end{aligned} \quad (18)$$

Because the basis functions are orthonormal on the 3D space (Eq. 7), by inserting Eq. 16 and 17 into Eq. 18, it gives $\mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] = \Gamma^p(g)^T \mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] \Gamma^{p'}(g)$ for all $g \in G$. Therefore, since the irrep is orthogonal $(\Gamma^p(g))^{-1} = \Gamma^p(g)^T$, it follows that

$$\Gamma^p(g) \mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] = \mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] \Gamma^{p'}(g) \quad (19)$$

for all $g \in G$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{3 \times 3}$, $p, p' \in \{1, \dots, N_{\text{rep}}\}$, $\zeta \in \{1, \dots, N_{p;\zeta}\}$, and $\zeta' \in \{1, \dots, N_{p';\zeta'}\}$. Since $\Gamma^p(g) \in \mathbb{R}^{d_p \times d_p}$ and $\Gamma^{p'}(g) \in \mathbb{R}^{d_{p'} \times d_{p'}}$, $\mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T]$ is a $d_p \times d_{p'}$ matrix. Eq. 19 has substantial structure because $\Gamma^p(g)$ and $\Gamma^{p'}(g)$ both for all $g \in G$ are irrep matrices. Schur's Lemma [12, Theorem I and II, Section 4.5, p. 80] implies that

$$\mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T] = \begin{cases} v_p((\zeta, \zeta')) I_{d_p}, & p = p' \\ \mathbf{0}_{d_p, d_{p'}}, & \text{otherwise} \end{cases} \quad (20)$$

for some number $v_p((\zeta, \zeta'))$ depending on p, ζ and ζ' , where $\mathbf{0}_{i,j}$ is the $i \times j$ zero matrix. Note that the solution of Eq. 19 for $\mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T]$ does not depend on ζ and ζ' except that unspecified degrees of freedom in the solution could depend on ζ and ζ' .

Suppose that the indices vary from the slowest to the fastest in the order p, ζ , or more specifically, in the order p, l, n , and q . Then p is constant over sequential sets of $d_p N_{p;\zeta}$ rows and columns in covariance matrix V . The matrix $\mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T]$ constructs the (p, p') th block of the covariance matrix V , denoted by $V_{p;p'}$. Precisely, $\mathbb{E}[c_{p;\zeta} c_{p';\zeta'}^T]$ is the (ζ, ζ') th sub-block of the matrix $V_{p;p'}$. Therefore, implied by Eq. 20, the covariance matrix V is a diagonal block matrix with five blocks, corresponding to the five values of p , and off diagonal blocks are zeros, i.e., $V_{p;p'} = \mathbf{0}_{(d_p N_{p;\zeta}), (d_{p'} N_{p';\zeta'})}$. The p th block itself ($V_{p;p}$) is a Kronecker product of the identity matrix I_{d_p} and a $N_{p;\zeta} \times N_{p;\zeta}$ matrix V_p , which has the (ζ, ζ') th $(\zeta, \zeta' \in \{1, \dots, N_{p;\zeta}\})$ entry to be $v_p(\zeta, \zeta')$. More explicitly,

$$\begin{aligned} V_{p;p} &= V_p \otimes I_{d_p} \\ &= \begin{bmatrix} v_p(1,1) & v_p(1,2) & \dots & v_p(1, N_{p;\zeta}) \\ v_p(2,1) & v_p(2,2) & \dots & v_p(2, N_{p;\zeta}) \\ \vdots & \vdots & \ddots & \vdots \\ v_p(N_{p;\zeta},1) & v_p(N_{p;\zeta},2) & \dots & v_p(N_{p;\zeta}, N_{p;\zeta}) \end{bmatrix} \otimes I_{d_p} \\ &= \begin{bmatrix} v_p(1,1)I_{d_p} & v_p(1,2)I_{d_p} & \dots & v_p(1, N_{p;\zeta})I_{d_p} \\ v_p(2,1)I_{d_p} & v_p(2,2)I_{d_p} & \dots & v_p(2, N_{p;\zeta})I_{d_p} \\ \vdots & \vdots & \ddots & \vdots \\ v_p(N_{p;\zeta},1)I_{d_p} & v_p(N_{p;\zeta},2)I_{d_p} & \dots & v_p(N_{p;\zeta}, N_{p;\zeta})I_{d_p} \end{bmatrix}. \end{aligned} \quad (21)$$

The covariance matrix V constructed from all of the $V_{p;p}$ blocks must be a real-valued positive semidefinite matrix, which requires each of the blocks is a real-valued positive semidefinite matrix. In order to make Eq. 21 be real-valued and positive semidefinite, it is necessary that $v_p(\zeta, \zeta') \in \mathbb{R}^+ \cup \{0\}$ for all $\zeta, \zeta' \in \{1, \dots, N_{\text{rep}}\}$ and $p \in \{1, \dots, N_{\text{rep}}\}$.

IV. HETEROGENEOUS RECONSTRUCTION WITH SYMMETRICAL STATISTICS (HRSS)

The 3D reconstruction process of heterogeneous virus particles which realize the symmetrical statistics is described in this section. In particular, the first two order statistics of the weights \bar{c} and V are solved by a constrained maximum likelihood estimator.

Suppose that a image stack includes N_y numbers of 2D projection images. Through the projection slice theorem, the model for the 3D structure of the virus particle (Eq. 4) can be linearly transformed into a model for the 2D projection images, specifically [3],

$$\mathbf{y}_i = L(z_i) \mathbf{c}_i + \mathbf{w}_i, \mathbf{w}_i \sim N(0, Q) \quad (22)$$

where \mathbf{y}_i is the i th 2D projection image in the reciprocal space, z_i is a vector of nuisance parameters which include the projection orientation of the i th image and the projection location of the particle in the i th image, $\mathbf{c}_i \in \mathbb{R}^{N_\tau}$ is the vector of weights of the i th image², $L(z_i)$ is the transformation matrix from weights to the i th reciprocal-space image which is transformed from the real-valued basis functions by projection slice theorem, and \mathbf{w}_i is the white Gaussian noise of the i th image which has zero mean and a covariance Q . Based on our assumption, the vector of weights is Gaussian distributed, specifically, $\mathbf{c}_i \sim N(\bar{\mathbf{c}}, V)$, where $\bar{\mathbf{c}}$ and V are the parameters to be determined.

²The weights \mathbf{c}_i in Eq. 22 are identical to the weights in the 3D model (Eq. 4).

We use a maximum likelihood estimator, as is described in [3], to determine \bar{c} and V . The maximum likelihood estimator is computed by an expectation-maximization algorithm with the nuisance parameters $\{z_i\}_{i=1}^{N_y}$. Three quantities are updated, the *a priori* distribution of the ensemble of virus particle, the mean \bar{c} and the covariance V of the weights. In our problem, the two quantities, \bar{c} and V , are updated alternatively in the expectation-maximization algorithm.

The previous algorithm HRSI [2], [14], which only used a subset of the proposed angular basis functions (just $p = 1$), used Matlab's `fmincon` (option "trust-region-reflective") with symbolic cost, gradient of the cost, and Hessian of the cost. We desired to modify the software to include the full set of basis functions described in Section II. The constraint on the mean vector (Eq. 15) requires its entries to be zero if $p \neq 1$. The constraints on the covariance matrix V (Eq. 21) require certain matrix elements to be equal. To implement the above constraints, the method of computing the gradient and Hessian of the cost is modified based on the chain rule: if f is the cost then

$$\frac{\partial f}{\partial \zeta_{p;l,n,q}} = \sum_{j=1}^{d_p} \frac{\partial f}{\partial \nu_{p,j;l,n,q}},$$

$$\frac{\partial^2 f}{\partial \zeta_{p_1;l_1,n_1,q_1} \partial \zeta_{p_2;l_2,n_2,q_2}} = \sum_{j_1=1}^{d_{p_1}} \sum_{j_2=1}^{d_{p_2}} \frac{\partial^2 f}{\partial \nu_{p_1,j_1;l_1,n_1,q_1} \partial \nu_{p_2,j_2;l_2,n_2,q_2}}.$$

These equations compute the necessary gradient and Hessian in terms of larger vectors and matrices which are then reduced in size. While this approach fits the software easily, more efficient approaches may be possible.

V. RECONSTRUCTION RESULTS

The performance of the proposed algorithm, HRSS, is evaluated in this section. The reconstruction process has been performed both on a simulated data (Section V-B) and on the experimental data of a Virus Like Particle (VLP) derived from bacteriophage HK97 Prohead I^{pro} (Section V-C). The VLP is essentially the bacteriophage minus the bacteriophage's tail leaving only the icosahedrally symmetric capsid. The average outer radius of the capsid is 254 Å.

A. Data processing

Two sets of 1200 cryo-EM images were separately processed. Each image containing one bacteriophage HK97 Prohead I^{pro} particle, was randomly selected from a larger stack. Each 2-D cryo-EM image measuring 200×200 pixels with a pixel size of 2.76 Å. Due to limitations of our computer hardware (computational speed and memory size), it is not practical to process the entire stack. The image selection algorithm [2] guarantees on non-overlapping sets of images. For each data set, following the preprocessing procedure in [2], [3], the 1200 cryo-EM images were transformed into 1200 2-D images in real and reciprocal space, and the reciprocal space images are the input to reconstruction algorithms.

The reconstruction is achieved in the following two steps.

- (a) The mean vector of the weights, denoted by \bar{c}_0 , is first computed by the *homogeneous reconstruction* algorithm (HRII) described in [3] which has $V_0 = 0$.

- (b) These results \bar{c}_0 and V_0 are used as the initial condition for the *heterogeneous reconstruction* (HRSI or HRSS) in the second step to compute the final results for \bar{c} and V .

Both HRSI and HRSS are tested. When testing on the existing heterogeneous reconstruction model (HRSI), due to the computing limitation, basis functions with l up to 55 and q up to 20 were employed in both Step (a) (HRII) and Step (b) (HRSI). The number of parameters to be estimated in Step (b) for either \bar{c} or V is 1060. When testing on the proposed heterogeneous reconstruction model (HRSS) which use the complete basis functions, we aim to achieve a similar scale of parameters to be estimated. Basis functions with l up to 10 and q up to 20 were employed in both Step (a) (HRII) and Step (b) (HRSS). The number of parameters to be estimated in Step (b) for \bar{c} or V is 2020.

B. Performances on the simulated data

The simulated 2D projection images in reciprocal space are generated through the same reconstruction forward model (Eq.22). The input of the simulator is the HRSS reconstruction results \bar{c} and V , which are called the "ground truth" in the stimulation study. The parameters to be determined ahead of the simulation process include the resolution (e.g., the pixel size in Å and the total number of pixels) and the signal-to-noise ratio (SNR) of the image. The method of generating the simulated 2D images is described in Algorithm 1. Let the

Algorithm 1 Generation of N simulated 2D images.

Data: The HRSS results \bar{c} and V

Result: a stack of 2D projection simulated images $\{y_i\}_{i=1}^N$

```

1 for  $i = 1; i \leq N; i++$  do
2   1. generate  $c_i \sim N(\bar{c}, V)$  and pick  $z_i$ 
3   2. compute the 2D image:
      (reciprocal space)  $y_i = L(z_i)c_i$ 
      (real space)  $y_i \rightarrow Real(y_i)$  [3]
4   3. construct the white noise:
      compute sample variances of  $Real(y_i)$ , denoted by  $s^2$ 
      generate  $w_i \sim N(0, Q)$ , where  $Q = \frac{s^2}{SNR}I$ 
5   4. generate simulated images (Eq.22):
      (reciprocal space)  $y_i = L(z_i)c_i + w_i$ 
      (real space)  $y_i \rightarrow Real(y_i)$  [3]
6 end
```

image have 100×100 pixels with a pixel size of 5.52 Å and let $SNR = 0.5$. A stack of 1200 simulated 2D projection images of HK97 is generated. Examples of the simulated real space 2D images of HK97 are shown in Fig. 3. With this input, the two-step reconstruction with the use of HRSS is performed. By using the same parameter settings in the reconstruction with the experimental images as the input, basis functions with l up to 10 and q up to 20 were employed in both Step (a) (HRII) and Step (b) (HRSS).

The standard measure of performance in structural biology is the Fourier shell correlation (FSC) ([15], Eq. 2; [16], Eq. 17; [17], p. 879) between two structures (the two estimates of mean electron scattering intensity) computed from the disjoint stacks of images. The final reconstruction (\bar{c} and V)

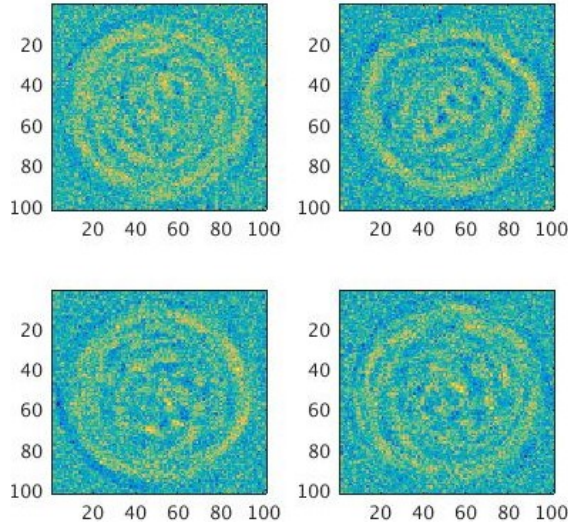


Fig. 3: Four simulated real space 2D images of HK97.

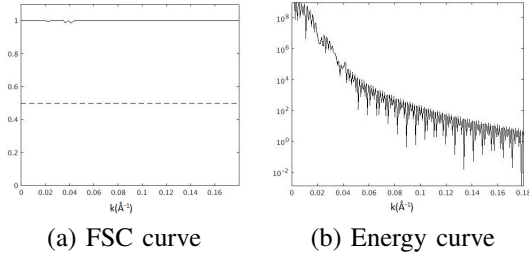


Fig. 4: FSC and energy curves between the “ground truth” and the reconstruction results of the simulated data.

is compared with the “ground truth”. Their FSC curve and energy curve (the denominators of FSC calculation) are shown in Fig. 4. The results show that the Fourier Shell Correlation (FSC) between the reconstruction result and the “ground truth” are almost 1 for more than the range $k \leq 0.186 \text{ \AA}^{-1}$ and therefore the structures are essentially identical for more than the range $k \leq 0.186 \text{ \AA}^{-1}$. The energy in both structures drops around 5 at 0.186 \AA^{-1} , which is greater than 0. Therefore, the FSC analysis in the range $k \leq 0.186 \text{ \AA}^{-1}$ is meaningful.

C. Performances on the experimental data

From the mean vector \bar{c} and the covariance matrix V , it is possible to compute the mean function [3, Eq. 16], $\bar{\rho}(\mathbf{x}) = E[\rho(\mathbf{x})]$, and covariance function [3, Eq. 18], $C_\rho(\mathbf{x}_1, \mathbf{x}_2) = E[(\rho(\mathbf{x}_1) - \bar{\rho}(\mathbf{x}_1))(\rho(\mathbf{x}_2) - \bar{\rho}(\mathbf{x}_2))]$, of the electron scattering intensity $\rho(\mathbf{x})$ of the particle. We often visualize $s_\rho(\mathbf{x}) = \sqrt{C_\rho(\mathbf{x}, \mathbf{x})}$ which has the same units as $\bar{\rho}(\mathbf{x})$. Through Eq. 9, the mean vectors \bar{c} computed by HRSI and HRSS in Step (b) were used to compute two mean functions $\bar{\rho}(\mathbf{x})$. Similarly, the two covariance matrices V computed by HRSI and HRSS were used to compute two standard deviation functions $s_\rho(\mathbf{x})$. We computed FSC between the reconstructions of the two data sets for both HRSI and HRSS, which are shown in Fig. 5. In the case of HRSI (Fig. 5 (a)), the resolution is determined by the fact that the FSC curve crosses the 0.5 level at 0.061 \AA^{-1} , which gives a resolution of 16.442 \AA . In the case of

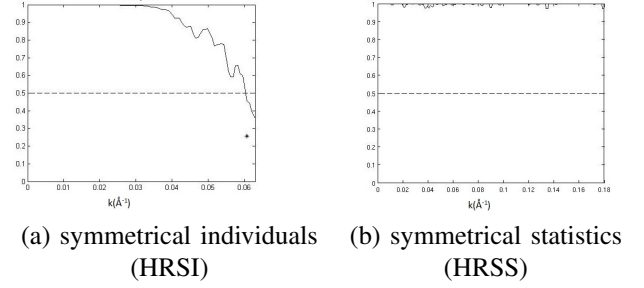
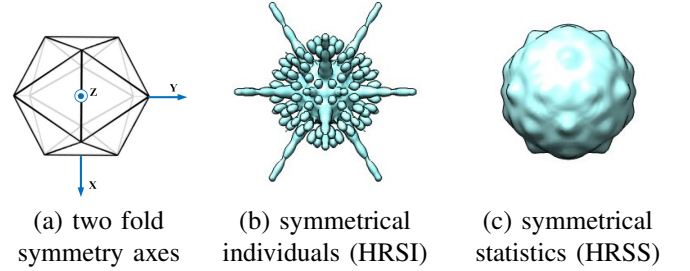


Fig. 5: FSC curves between the reconstructions of the two experimental data sets for HRSI (a) and HRSS (b).

Fig. 6: 3-D reconstructions of the standard deviation s_ρ for HK97 Prohead I^{Pro+}. The shape is a surface of constant intensity (0.0038) of the standard deviation $s_\rho(\mathbf{x})$, which is visualized by UCSF Chimera [18].

HRSS (Fig. 5 (b)), the FSC between the two reconstructions are almost 1 for more than the range $k \leq 0.186 \text{ \AA}^{-1}$ and therefore the structures are essentially identical.

Fig. 6 (b) and (c) show the standard deviation $s_\rho(\mathbf{x})$ computed by HRSI and HRSS, respectively. In Fig. 7, $\bar{\rho}(\mathbf{x})$ computed by HRSI and the two heterogeneous solutions $s_\rho(\mathbf{x})$ computed by HRSI and HRSS are jointly visualized by computing a 3-D surface of constant value of $\bar{\rho}(\mathbf{x})$ and coloring the surface by $s_\rho(\mathbf{x})$. The cubes of $\bar{\rho}$ and s_ρ in Fig. 6–7 have the same orientation as shown in Fig. 6 (a), in which the x-z plane contains the two fold symmetry of the icosahedron. We call Fig. 6 (a) as the two fold axes.

Fig. 8 shows six different cross sections of the standard deviation function $s_\rho(\mathbf{x})$ in three different axes the two fold, the three fold and the five fold symmetry axes. The two fold symmetry, the three fold symmetry and the five fold symmetry of the icosahedron are sitting along the Z axis, as shown in the top row of Fig. 8, where the center of the s_ρ cube has the coordinate (300, 300, 300).

As shown in Fig. 6 (b) and Fig. 7 (a), the existing algorithm HRSI estimates a standard deviation function that is organized in radially-directed rays, which has no biological explanation. Specifically, HRSI introduced artificial spikes laying on the symmetry axes of the icosahedron. In fact, this artificial pattern in standard deviation or variance analysis of spherical virus particles has been recognized as a well-known problem of all existing reconstruction algorithms [19, p. 173], including the standard algorithms EMAN2 [20] and SPIDER [21].

In contrast, HRSS gives a standard deviation function that is

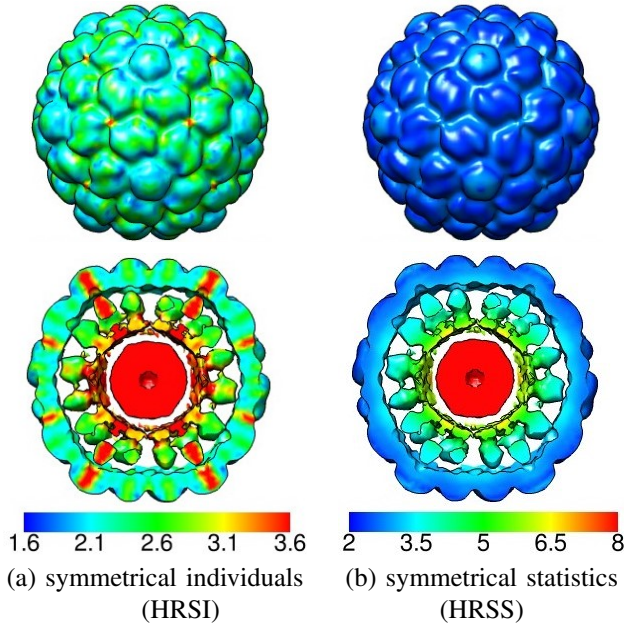


Fig. 7: 3-D reconstructions of $\bar{\rho}$ for HK97 Prohead I⁺pro (different colormap). The shape is a surface of constant intensity (5×10^{-4}) of $\bar{\rho}(\mathbf{x})$ colored by the standard deviation $s_{\rho}(\mathbf{x})$, which is visualized by UCSF Chimera [18]. The HRSI and HRSS reconstructions use different color maps. All markings are scaled by 10^{-3} .

organized in *annular* structures as shown in Fig. 6 (c) and more obviously in Fig. 7 (b). Such an annular structure obtained by HRSS matches the physical organization of the particle, for instances, the outer protein shell and the inner core of nucleic acid. This makes more biological sense than the estimates of HRSI and other reconstruction algorithms. Such a distinction can be further seen in Fig. 8, in which Fig. 8 (a) has light blue radial lines, whereas the Fig. 8 (b) shows the detailed annular structure of the virus. In the previous analyses with the use of HRSI, features like high values located along radially directed lines averaged out leaving interpretable information, but as we try to increase the spatial detail of our interpretation, such behavior is difficult to understand.

VI. CONCLUSION

In this paper, we proposed a novel approach HRSS for 3-D reconstruction of nanoscale virus particles from noisy 2-D projection images. Realistic assumption on the heterogeneous virus particles requires sophisticated tools. HRSS incorporates the complete basis functions to realize the sophisticated situation of symmetrical statistics, which eliminates the long-standing distortions in heterogeneity calculations associated with symmetry axes, and provides estimates that make more biological sense than the previous estimates. This implies that the variability in the particles is such that individual particles lack symmetry and only the first and second order *statistics* obey the symmetry. This is natural for such large objects as viruses ($10^2 - 10^3 \text{ \AA}$, 10 Megadalton) in the absence of the

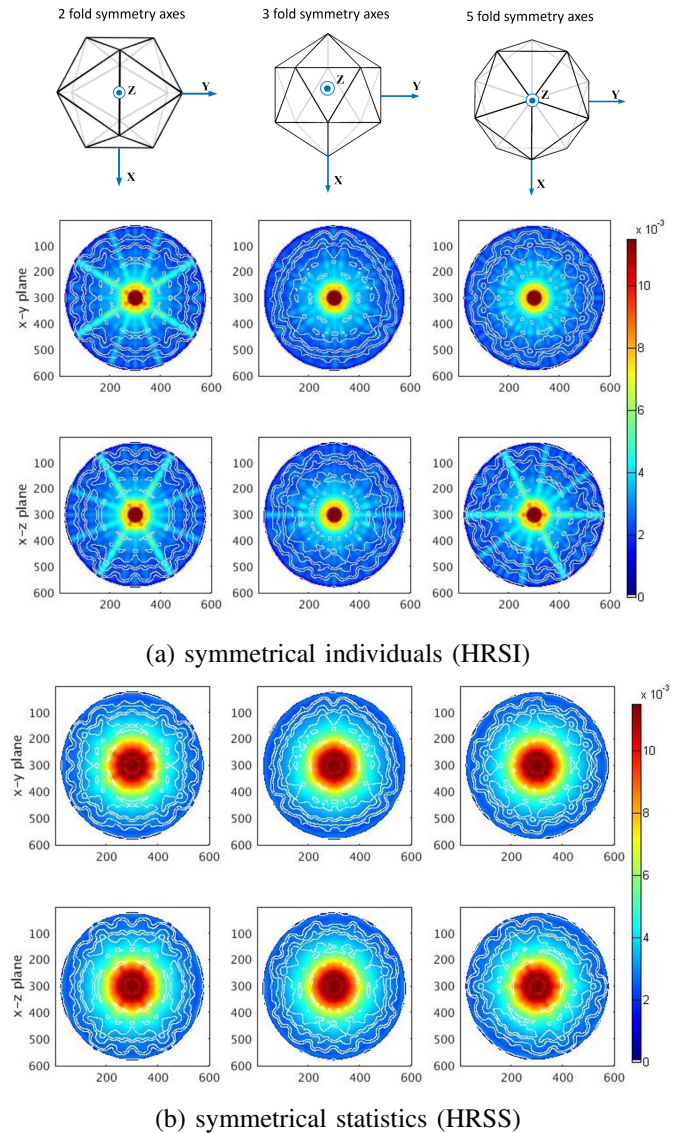


Fig. 8: Cross section of the standard deviation function $s_{\rho}(\mathbf{x})$ for HK97 Prohead I⁺pro using both HRSI and HRSS and displayed with a common color map. The white lines depict the contour levels of the virus particles. The geometrical orientation of the cubes are shown in the top row. (a) shows the reconstruction results based on symmetrical individuals, i.e., HRSI, whereas (b) shows the reconstruction results based on symmetrical statistics, i.e., HRSS. For each orientation, cross sections in x-y and x-z planes are displayed. The center of the s_{ρ} cube has the coordinate (300, 300, 300).

additional geometric constraints that occur in the crystal of an x-ray crystallography experiments.

VII. ACKNOWLEDGMENTS

We are grateful to Professors John E. Johnson (The Scripps Research Institute) and David Veesler (University of Washington) for the HK97 data and helpful discussions.

APPENDIX A DERIVATION OF EQ. 7

The orthonormal condition of angular basis functions $I_{p;l,n}$ (Property 2 in Section II) explicitly is [10]

$$\int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} I_{p;l,n}^T(\mathbf{x}/x) I_{p';l',n'}(\mathbf{x}/x) d\theta d\phi = I_{d_p} \delta_{p,p'} \delta_{l,l'} \delta_{n,n'}, \quad (23)$$

where \mathbf{x}/x is shorthand for (θ, ϕ) . By Eq. 23 and Eq. 6 (the orthonormality of radial basis functions), the LHS of 7 can be expressed as

$$\begin{aligned} & \int_{\mathbf{x}} F_{p;\zeta}(\mathbf{x}) F_{p';\zeta'}^T(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbf{x}} (I_{p;l,n}(\mathbf{x}/x) h_{l,q}(x)) (I_{p';l',n'}(\mathbf{x}/x) h_{l',q'}(x))^T d\mathbf{x} \\ &= \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} \int_{x=0}^{\infty} (I_{p;l,n}(\mathbf{x}/x) h_{l,q}(x)) (I_{p';l',n'}(\mathbf{x}/x) h_{l',q'}(x))^T dx d\phi d\theta \\ &= \int_{\theta=0}^{\pi} I_{p;l,n}(\mathbf{x}/x) \left(\int_{\phi=0}^{2\pi} h_{l,q}(x) h_{l',q'}^T(x) dx \right) I_{p';l',n'}^T(\mathbf{x}/x) d\phi d\theta \\ &= (\delta_{l,l'} \delta_{q,q'}) \int_{\theta=0}^{\pi} I_{p;l,n}(\mathbf{x}/x) I_{p';l',n'}^T(\mathbf{x}/x) d\phi d\theta \\ &= I_{d_p} \delta_{p,p'} \delta_{l,l'} \delta_{n,n'} \delta_{q,q'} \\ &= I_{d_p} \delta_{p,p'} \delta_{\zeta,\zeta'} \end{aligned} \quad (24)$$

ACKNOWLEDGMENT

We are grateful to Professors John E. Johnson (The Scripps Research Institute) and David Veesler (University of Washington) for the HK97 data and helpful discussions.

REFERENCES

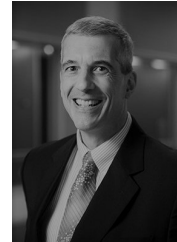
- [1] G. J. Jensen, Ed., *Cryo-EM, Parts A–C*, ser. Methods in Enzymology. Elsevier Inc., 2010, vol. 481–483.
- [2] Q. Wang, T. Matsui, T. Domitrovic, Y. Zheng, P. C. Doerschuk, and J. E. Johnson, “Dynamics in cryo EM reconstructions visualized with maximum-likelihood derived variance maps,” *J. Struct. Biol.*, vol. 181, no. 3, pp. 195–206, Mar. 2013.
- [3] Y. Zheng, Q. Wang, and P. C. Doerschuk, “3-D reconstruction of the statistics of heterogeneous objects from a collection of one projection image of each object,” *J. Opt. Soc. Am. A*, vol. 29, no. 6, pp. 959–970, Jun. 2012.
- [4] S. Baron, *Epidemiology–Medical Microbiology*. University of Texas Medical Branch at Galveston, 1996.
- [5] P. C. Doerschuk and J. E. Johnson, “*Ab initio* reconstruction and experimental design for cryo electron microscopy,” *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1714–1729, Aug. 2000, <http://dx.doi.org/10.1109/18.857786>.
- [6] Z. Yin, Y. Zheng, P. C. Doerschuk, P. Natarajan, and J. E. Johnson, “A statistical approach to computer processing of cryo electron microscope images: Virion classification and 3-D reconstruction,” *J. Struct. Biol.*, vol. 144, no. 1/2, pp. 24–50, 2003, <http://www.ncbi.nlm.nih.gov/pubmed/14643207>.
- [7] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J.-M. Carazo, “Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization,” *Nature Methods*, vol. 4, no. 1, pp. 27–29, Jan. 2007.
- [8] J. Lee, P. C. Doerschuk, and J. E. Johnson, “Exact reduced-complexity maximum likelihood reconstruction of multiple 3-D objects from unlabeled unoriented 2-D projections and electron microscopy of viruses,” *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2865–2878, Nov. 2007, <http://www.ncbi.nlm.nih.gov/pubmed/18092587>.
- [9] R. A. Redner and H. F. Walker, “Mixture densities, maximum likelihood and the EM algorithm,” *SIAM Review*, vol. 26, no. 2, pp. 195–239, Apr. 1984.
- [10] N. Xu and P. C. Doerschuk, “Computation of real basis functions for the 3-d rotational polyhedral groups t , o , and i ,” *Submitted to SIAM Journal of Mathematical Analysis*, 2016.

- [11] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, Eds., *NIST Handbook of Mathematical Functions*. Cambridge, UK: Cambridge University Press, 2010. [Online]. Available: <http://dlmf.nist.gov>
- [12] J. F. Cornwell, *Group Theory in Physics*. London: Academic Press, 1984, vol. 1.
- [13] Y. Zheng and P. C. Doerschuk, “3D image reconstruction from averaged Fourier transform magnitude by parameter estimation,” *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1561–1570, Nov. 1998.
- [14] J. Tang, B. M. Kearney, Q. Wang, P. C. Doerschuk, T. S. Baker, and J. E. Johnson, “Dynamic and geometric analyses of *Nudaurelia capensis* ω virus maturation reveal the energy landscape of particle transitions,” *J. Molecular Recognition*, vol. 27, no. 4, pp. 230–237, 10 February 2014, <http://www.ncbi.nlm.nih.gov/pubmed/24591180>.
- [15] M. van Heel, “Similarity measures between images,” *Ultramicroscopy*, vol. 21, pp. 95–100, 1987.
- [16] G. Harauz and M. van Heel, “Exact filters for general geometry three dimensional reconstruction,” *Optik*, vol. 73, no. 4, pp. 146–156, 1986.
- [17] T. S. Baker, N. H. Olson, and S. D. Fuller, “Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs,” *Microbiology and Molecular Biology Reviews*, vol. 63, no. 4, pp. 862–922, Dec. 1999.
- [18] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, “UCSF Chimera—A visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.
- [19] S. Ludtke, “Chapter seven-single-particle refinement and variability analysis in eman2. 1,” *Methods in Enzymology*, vol. 579, pp. 159–189, 2016.
- [20] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, “Eman2: an extensible image processing suite for electron microscopy,” *Journal of structural biology*, vol. 157, no. 1, pp. 38–46, 2007.
- [21] T. R. Shaikh, H. Gao, W. T. Baxter, F. J. Asturias, N. Boisset, A. Leith, and J. Frank, “Spider image processing for single-particle reconstruction of biological macromolecules from electron micrographs,” *Nature protocols*, vol. 3, no. 12, pp. 1941–1974, 2008.



Nan Xu is pursuing her Ph.D. degree in School of Electrical and Computer Engineering with a minor in Applied Mathematics and a minor in Cognitive Neuroscience at Cornell University. Nan received her Master of Science degree in May 2015. Before joining Cornell, she received double Bachelor’s degrees in Electrical and Computer Engineering (BS) and Mathematics (BA) with a minor in Music at the University of Rochester in Rochester NY. Nan’s current research interests are in statistical modeling and inference in biological data. She is investigating

three primary applications: brain network estimation, realistic fMRI data simulation, and reconstruction of the statistical characteristics of ensembles of heterogeneous virus particles.



Peter C Doerschuk is a professor in the School of Electrical and Computer Engineering and Nancy E. and Peter C. Meing School of Biomedical Engineering at Cornell University. His research concerns biological and medical systems from the view point of computational nonlinear stochastic systems. In particular, he has contributed to computational inverse problems for biophysics, statistical image processing, and biomedical and speech signal processing. In each of these areas, important goals of his work have been to incorporate accurate physical models while

at the same time developing computationally practical algorithms implemented in high-performance software systems.

Before joining Cornell University in July 2006, Peter Doerschuk was on the faculty of Purdue University in Electrical and Computer Engineering and Biomedical Engineering. He received BS, MS, and Ph.D. degrees in Electrical Engineering from MIT and an M.D. degree from Harvard Medical School. After post-graduate training at Brigham and Womens’ Hospital he held a post-doctoral appointment at the Laboratory for Information and Decision Systems (MIT) before joining Purdue.